CMPUT 566 Basic Mini-Project Report

University of Alberta
Department of Computing Science

Zirui Liu
1573088

Introduction

The basic goal of this project is to use three different machine learning algorithms on a hear disease prediction dataset. The dataset is obtained from the Kaggle platform where machine learning engineers share the resources. The dataset contains 4239 samples and there are 15 features used for predicting weather the patient has heart disease or not. This is a binary classification problem because the value of heart disease is either 0 or 1. The features include Gender, ageeducation, currentSmoker, cigsPerDay,BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP,BMI, heartrate,glucose.

	Gender	age	education	currentSmoker	cigsPerDay	BPMeds	\
1	Female	46	primaryschool	0	0.0	0.0	
2	Male	48	uneducated	1	20.0	0.0	
3	Female	61	graduate	1	30.0	0.0	
4	Female	46	graduate	1	23.0	0.0	
5	Female	43	primaryschool	0	0.0	0.0	
..	...	...	...	...	...	...	
95	Female	65	graduate	0	0.0	0.0	
96	Female	63	postgraduate	1	20.0	0.0	
97	Female	40	primaryschool	0	0.0	0.0	
98	Female	56	uneducated	0	0.0	0.0	
99	Female	56	uneducated	1	15.0	0.0	

Machine Learning Algorithms

There are three machine learning algorithms used for solving this problem.

Support Vector Machine

The first algorithm is SVM. Since the input has 15 features and the first I did was to convert the heart_stroke column which is the output form YES/NO to 1/0 since this is a binary classification problem. Similarly, I did the same thing with prevalentStroke colum which is one of the features, Since I think all of the features are necessary to predict if the person have heart disease so I didn’t dropout any input values. The first step in the model development process is handling missing values in the 'glucose' column. The missing values are imputed using the mean of the available data, ensuring a continuous and complete feature for analysis. Additionally, categorical variables ('Gender' and 'education') are converted into numerical format using one-hot encoding, while binary categorical variables ('prevalentStroke' and 'Heart_ stroke') are encoded

using label encoding.

The input dataset was split into train, test, validation set in the ratio of 0.7, 0.15, and 0.15 respectively.

Hyperparameter tuning is a crucial step in machine learning model development, as it involves optimizing the parameters that are not learned from the training data but significantly impact the model's performance. In this report, we focus on hyperparameter tuning for a Support Vector Machine (SVM) using the GridSearchCV technique. The objective is to identify the best combination of hyperparameters that maximize the model's predictive performance. GridSearchCV Parameters:

The hyperparameter grid used for the GridSearchCV is defined as follows:

- **C (Regularization Parameter):** [0.001, 0.01, 0.1, 1, 10, 100]
- **Kernel Function:** ['linear', 'rbf', 'poly']
- **Gamma:** ['scale', 'auto']

GridSearchCV is a technique that exhaustively searches the specified hyperparameter grid using cross-validation. In this case, a 5-fold cross-validation strategy is employed, dividing the training dataset into five subsets. The SVM model is trained and evaluated using different hyperparameter combinations, and the best combination is selected based on the average performance across these folds.

The result of **svm** is shown below.

Best Hyperparameters: {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
Validation Accuracy: 0.8444444444444444
Validation Classification Report:
Testing Accuracy: 0.782608695652174

Logical Regression

The second machine learning algorithm is logical regression, which can be used for classifying binary problem, the data processing steps are the same the SVM. Logistic regression is chosen as the predictive model for its simplicity and interpretability. Hyperparameter tuning is performed using GridSearchCV, which systematically searches a predefined hyperparameter grid to identify the optimal configuration. The hyperparameter tuned is the regularization parameter 'C,' with values ranging from 0.001 to 100. The result of this algorithm is shown below.

Best Hyperparameters: {'C': 10}
Validation Accuracy: 0.8710691823899371
Testing Accuracy: 0.8537735849056604

Linear Regression

The last algorithm used was linear regression. In this analysis, we employed a linear regression model to predict the risk of heart stroke based on a set of health-related features. The dataset underwent preprocessing steps, including label encoding for binary categorical variables, splitting into training, validation, and test sets, and imputing missing values with the mean. The linear regression model was trained on the combined training and validation sets and evaluated on both the validation and test sets using key metrics such as Mean Squared Error (MSE) and R-squared.

The result is shown below.

Mean Squared Error on Validation Set: 0.8906541002479454
R-squared on Validation Set: 0.9000809740762528
Mean Squared Error on Test Set: 0.8810751154678625
R-squared on Test Set: 0.9123860119435455

Conclusion

As can be seen from the result, the testing accuracy of linear regression is the highest among these three algorithms. The matrix is based on validation accuracy and testing accuracy only.